

Modification of the MDR-EFE method for stratified samples

Alexander Bulinski¹, Alexey Kozhevin²

The MDR-EFE method of performing identification of relevant factors within a given collection X_1, \dots, X_n is developed for stratified samples in the case of binary response variable Y . We establish a criterion of strong consistency of estimates (involving K -cross-validation procedure and penalty) for a specified prediction error function. The cost approach is proposed to compare experiments with random and nonrandom number of observations. Analytic results and simulations demonstrate advantages of the method introduced for stratified samples over that employed for i.i.d. learning sample.

Keywords: Feature selection; MDR method; Error function estimation; Cross-validation; Stratified sample; Cost approach.

1 Introduction

The research direction combining probability, statistics and machine learning for analysis of mathematical problems of feature selection is vastly represented in literature along with various applications of this theory. Quite a number of powerful methods were developed for different models in the course of such investigations. Several new variable selection procedures have emerged during the last 20 years. One can refer, e.g., to the following books [1] – [3], [13] – [15], [19]. Note that many exhaustive, stochastic and heuristic methods to detect epistasis (in genetics) are considered in [6], [16] and [20].

In the paper by M.Ritchie et al. [18] the *multifactor dimensionality reduction* (MDR) method was proposed to identify the relevant (in a sense) factors having influence on a binary response variable. The review [12] demonstrates great popularity of this method. Some 800 papers published between 2001 and 2014 were devoted to extensions, modifications and applications of the general idea suggested in [18]. The goal of the present paper is to extend the dimensionality reduction of factors to stratified samples framework. Here we generalize the approach developed in [4]–[9] for i.i.d. observations.

Recall some notation. Let X_1, \dots, X_n be a collection of random features (or factors) and Y be a response variable depending on $X := (X_1, \dots, X_n)$. We assume that all random elements under consideration are defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Suppose that X_i takes values in a finite set \mathbb{X}_i , $i = 1, \dots, n$. Thus $X: \Omega \rightarrow \mathbb{X}$ where $\mathbb{X} := \mathbb{X}_1 \times \dots \times \mathbb{X}_n$. We consider Y with values in $\{-1, 1\}$ (one uses also the set $\{0, 1\}$). There are important models in medicine and biology where Y characterizes the health state of a patient (for example $Y = 1$ means the *case*, i.e. a patient is sick, and $Y = -1$ corresponds to the *control*, that is a person is healthy) and X comprises both genetic and nongenetic factors. The challenging problem is to predict the risk of certain complex disease on account of the data X and also

¹Address: Faculty of Mathematics and Mechanics of the Lomonosov Moscow State University, Moscow 119991, Russia. E-mail: bulinski@mech.math.msu.su

²Faculty of Mathematics and Mechanics of the Lomonosov Moscow State University. E-mail: kozhevin.alexey@gmail.com

to identify the collection $(X_{i_1}, \dots, X_{i_r})$ of relevant factors ($r < n$) which are responsible for the disease provoking.

For any $f: \mathbb{X} \rightarrow \{-1, 1\}$ the quality of the forecast of Y by means of $f(X)$ can be expressed by the following *Error function* (see [4], [7])

$$Err(f) = \mathbb{E}|Y - f(X)|\psi(Y)$$

where a *penalty function* $\psi: \{-1, 1\} \rightarrow \mathbb{R}_+$ is introduced to weight the importance of incorrect prediction of different values of Y . The trivial cases $\psi \equiv 0$, $Y \equiv 0$ or $Y \equiv 1$ are excluded. Clearly,

$$Err(f) = 2 \sum_{y \in \{-1, 1\}} \psi(y) \mathbb{P}(Y = y, f(X) \neq y). \quad (1)$$

All the *optimal functions* f_{opt} , i.e. f rendering minimum to $Err(f)$, were described in [4]. It is convenient to take the optimal function $f^*(x) = \mathbb{I}\{A\}(x) - \mathbb{I}\{\bar{A}\}(x)$, $x \in \mathbb{X}$, where

$$A = \left\{ x \in M : \mathbb{P}(Y = 1 | X = x) > \frac{\psi(-1)}{\psi(-1) + \psi(1)} \right\}, \quad (2)$$

$M = \{x \in \mathbb{X} : \mathbb{P}(X = x) > 0\}$, $\mathbb{I}\{A\}$ stands for indicator of a set A and \bar{A} means the complement of A . Note that any optimal function f_{opt} gives the same value to Error function as $Err(f^*)$. In [4] it is explained why the choice of

$$\psi(y) = \frac{1}{\mathbb{P}(Y = y)}, \quad y \in \{-1, 1\}, \quad (3)$$

considered in [22] is natural.

However the joint law of (X, Y) is unknown, and therefore $Err(f)$ is unknown (f^* is unknown as well). So it is reasonable to apply for inference statistical estimates of $Err(f)$ involving independent identically distributed (i.i.d) random vectors $(X^1, Y^1), \dots, (X^N, Y^N)$, having the same law as (X, Y) . In [4]–[9] it is shown how one can use such estimates to identify the collection of relevant factors. We call such method MDR-EFE (*multifactor dimensionality reduction - error function estimation*). Moreover, in [5], [8] and [9] the asymptotic normality of introduced statistics was established and in two latter papers a nonbinary response was studied.

Now we concentrate on the following problem. Assume that the probability (of disease) $\mathbb{P}(Y = 1)$ is rather small. Then in the sample $(X^1, Y^1), \dots, (X^N, Y^N)$ for N not too large we could have too small observations amount with positive response variable value (equal to one) for sound conclusions. We discuss several scenarios to overcome this difficulty. Namely, we provide modifications of some previous results using stratification and also consider the random number of random observations.

The paper is organized as follows. After the brief Introduction (Section 1) in Section 2 we prove an auxiliary result concerning the law of observations having a given response value. Section 3 contains the main result. Here we provide a criterion of strong consistency of statistics which are the estimates of prediction error (of approximation of a response Y by means of a function of factors X_1, \dots, X_n) for stratified samples. We employ the cross-validation technique or, more precisely, the method of subsampling and averaging. Along with discussion of established result we consider an example showing its application. Then we demonstrate how one can apply the main result for feature selection. Section 4 is devoted to

the cost approach to experiments and the XOR-model (see [23]). Here we compare the MDR-EFE method for i.i.d. observations and the version of this method for stratified samples. The simulation results (Section 5) show that even for rather small samples the proposed version of the MDR-EFE method has visible advantages.

2 Auxiliary result

Let $(X, Y), (X^1, Y^1), (X^2, Y^2), \dots$ be a sequence of i.i.d. random vectors defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each $\omega \in \Omega$ we consider a sequence $Y^1(\omega), Y^2(\omega), \dots$ and pick all the indices $1 \leq j_{-1}^1(\omega) < j_{-1}^2(\omega) < \dots$ for which $Y^{j_{-1}^k(\omega)}(\omega) = -1$, $k \in \mathbb{N}$. In a similar way we will write all observations $Y^i(\omega)$ with values 1 as $\{Y^{j_1^m(\omega)}(\omega)\}_{m \in \mathbb{N}}$ where $1 \leq j_1^1(\omega) < j_1^2(\omega) < \dots$. Recall that for the Bernoulli trials with probability of success p the Negative binomial random variable $U_{r,p}$ is introduced as the number of successes needed to get r failures where $r \in \mathbb{N}$ (one writes $U_{r,p} \sim NB(r, p)$). Thus

$$\mathbb{P}(U_{r,p} = k) = \binom{k+r-1}{k} p^k (1-p)^r, \quad k = 0, 1, \dots$$

If we consider the events $\{Y^i = 1\}$ and $\{Y^i = -1\}$ as a success and failure, respectively (with probability $p = \mathbb{P}(Y = 1)$ of success), then j_{-1}^r has the same law as $U_{r,p} + r$. Therefore

$$\mathbb{P}(j_{-1}^r = m) = \begin{cases} \binom{m-1}{m-r} p^{m-r} (1-p)^r, & m = r, r+1, \dots, \\ 0, & m = 1, \dots, r-1. \end{cases} \quad (4)$$

For $r = 1$ we keep only the first line in (4) as in this case $\{1, \dots, r-1\} = \emptyset$. By similar reasons j_1^r is distributed as $U_{r,1-p} + r$ where $U_{r,1-p} \sim NB(r, 1-p)$. In other words j_1^r has the same law as $G_p^1 + \dots + G_p^r$ where G_p^1, \dots, G_p^r are independent random variables having the Geometric law with parameter p (i.e. $\mathbb{P}(G_p^1 = k) = p(1-p)^{k-1}$, $k = 1, 2, \dots$). Thus $\mathbb{E}j_1^r = \frac{r}{p}$ and $j_1^r < \infty$ a.s. for any $r \in \mathbb{N}$ (analogously $\mathbb{E}j_{-1}^r = \frac{r}{1-p}$ for each $r \in \mathbb{N}$). Note also that one can find different definitions of Negative binomial and Geometric laws, that is why we provided the explicit formulae.

Set $Z^k := X^{j_1^k}$ for each $k \in \mathbb{N}$. Introduce a collection \mathcal{B} of all subsets of \mathbb{X} . We use the following simple result.

Lemma 1 *For each $m \in \mathbb{N}$, the random variables Z^1, \dots, Z^m are independent and distributed as X given $Y = 1$ (we write $X|Y = 1$), i.e., for any $B \in \mathcal{B}$ and $k = 1, \dots, m$,*

$$\mathbb{P}(Z^k \in B) = \mathbb{P}(X \in B|Y = 1).$$

Proof. First of all we show that, for all $B_i \in \mathcal{B}$, $i = 1, \dots, m$,

$$\mathbb{P}(Z^1 \in B_1, \dots, Z^m \in B_m) = \mathbb{P}(Z^1 \in B_1) \dots \mathbb{P}(Z^m \in B_m). \quad (5)$$

By the total probability formula

$$\begin{aligned} & \mathbb{P}(Z^1 \in B_1, \dots, Z^m \in B_m) = \mathbb{P}(X^{j_1^1} \in B_1, \dots, X^{j_1^m} \in B_m) \\ &= \sum_{\substack{(k_1, \dots, k_m) \in \mathbb{N}^m: \\ k_1 < \dots < k_m}} \mathbb{P}(X^{j_1^1} \in B_1, \dots, X^{j_1^m} \in B_m, j_1^1 = k_1, \dots, j_1^m = k_m). \end{aligned}$$

Note that for arbitrary positive integers $k_1 < k_2 < \dots < k_m$

$$\{j_1^1 = k_1, \dots, j_1^m = k_m\} = \bigcap_{i=1}^m \{Y^{k_i} = 1\} \cap \bigcap_{r \in T_m} \{Y^r = -1\},$$

where $T_m = T_m(k_1, k_2, \dots, k_m) := \{1, \dots, k_m\} \setminus \{k_1, \dots, k_m\}$. Next, due to independence of the random vectors $(X^1, Y^1), (X^2, Y^2), \dots$, one has

$$\begin{aligned} & \mathbb{P}(Z^1 \in B_1, \dots, Z^m \in B_m) \\ &= \sum_{\substack{(k_1, \dots, k_m) \in \mathbb{N}^m: \\ k_1 < \dots < k_m}} \mathbb{P} \left(\{X^{k_1} \in B_1, \dots, X^{k_m} \in B_m, Y^{k_1} = 1, \dots, Y^{k_m} = 1\} \cap \left\{ \bigcap_{r \in T_m} \{Y^r = -1\} \right\} \right) \\ &= \sum_{\substack{(k_1, \dots, k_m) \in \mathbb{N}^m: \\ k_1 < \dots < k_m}} \mathbb{P}(X^{k_1} \in B_1, Y^{k_1} = 1) \dots \mathbb{P}(X^{k_m} \in B_m, Y^{k_m} = 1) \mathbb{P} \left(\bigcap_{r \in T_m} \{Y^r = -1\} \right) \\ &= \sum_{\substack{(k_1, \dots, k_m) \in \mathbb{N}^m: \\ k_1 < \dots < k_m}} \prod_{q=1}^m \frac{\mathbb{P}(X^{k_q} \in B_q, Y^{k_q} = 1)}{\mathbb{P}(Y^{k_q} = 1)} \mathbb{P} \left(\bigcap_{i=1}^m \{Y^{k_i} = 1\} \cap \bigcap_{r \in T_m} \{Y^r = -1\} \right) \\ &= \prod_{q=1}^m \mathbb{P}(X \in B_q | Y = 1) \sum_{\substack{(k_1, \dots, k_m) \in \mathbb{N}^m: \\ k_1 < \dots < k_m}} \mathbb{P}(j_1^1 = k_1, \dots, j_1^m = k_m) = \prod_{q=1}^m \mathbb{P}(X \in B_q | Y = 1). \quad (6) \end{aligned}$$

If we take $B_1 = \dots = B_{k-1} = B_{k+1} = \dots = B_m = \mathbb{X}$ in (6) then

$$\mathbb{P}(Z^k \in B_k) = \mathbb{P}(X \in B_k | Y = 1), \quad k = 1, \dots, m. \quad (7)$$

Hence, in view of (7) the random variables Z^1, \dots, Z^m are identically distributed and relation (5) holds. \square

Remark 1. Lemma 1 is also valid for i.i.d. vectors $(X, Y), (X^1, Y^1), \dots, (X^N, Y^N)$ when X takes values in any space \mathbf{S} endowed with some σ -algebra \mathcal{B} (X is measurable w.r.t. σ -algebras \mathcal{F} and \mathcal{B}). In the same way one can prove that $X^{j_{-1}^1}, X^{j_{-1}^2}, \dots$ are independent and identically distributed as $X|Y = -1$.

3 The main result and discussion

Let $(X^1, Y^1), (X^2, Y^2), \dots$ be i.i.d. observations having the same law as (X, Y) . Assume that we have a possibility to form a sample with given nonrandom numbers of cases and controls (N_1 and N_{-1} respectively). More exactly, we take

$$\zeta_{N_1}^1 := \{(X^{j_1^1}, 1), \dots, (X^{j_1^{N_1}}, 1)\}, \quad \zeta_{N_{-1}}^{-1} := \{(X^{j_{-1}^1}, -1), \dots, (X^{j_{-1}^{N_{-1}}}, -1)\}$$

where the random indices j_1^k and j_{-1}^k , $k \in \mathbb{N}$, were introduced in Section 2. Let $N := N_1 + N_{-1}$ be the size of our *stratified sample* $\zeta_N := \zeta_{N_1}^1 \cup \zeta_{N_{-1}}^{-1}$. In contrast to the i.i.d. sample $\xi_N := \{(X^1, Y^1), \dots, (X^N, Y^N)\}$ taken from the population with $\text{law}(X, Y)$ we have (according to Lemma 1) two subsamples $\zeta_{N_1}^1$ and $\zeta_{N_{-1}}^{-1}$ with laws $X|Y = 1$ and $X|Y = -1$, respectively.

Thus one cannot use the frequency estimates (e.g., of $P(Y = 1)$ or $P(X \in B, Y = 1)$ where $B \subset \mathbb{X}$) constructed by means of ζ_N . Further on we assume that $N_1 = \max\{[aN], 1\}$ and $N_{-1} = N - N_1$, here the parameter $a \in (0, 1)$, $N \in \mathbb{N}$ and $[\cdot]$ stands for the integer part of a number. Suppose that there are some estimators \hat{P}_N^y of $P(Y = y)$ such that

$$\hat{P}_N^y \rightarrow P(Y = y) \text{ a.s., } N \rightarrow \infty, \quad y \in \{-1, 1\}. \quad (8)$$

For instance we can assume that \hat{P}_N^y involve the data $\{(X^k, Y^k), 1 \leq k \leq \max\{j_1^{N_1}, j_{-1}^{N_{-1}}\}\}$. In this case the frequency estimates of the probabilities $P(Y = 1)$ or $P(X \in B, Y = 1)$, where $B \subset \mathbb{X}$, mentioned above are strongly consistent since $\max\{j_1^{N_1}, j_{-1}^{N_{-1}}\} \rightarrow \infty$ a.s. when $N \rightarrow \infty$. In Section 4 of the paper we discuss the advantages and disadvantages of employing the stratified samples. Introduce the vector $\hat{P}_N := (\hat{P}_N^{-1}, \hat{P}_N^1)$. Recall that we exclude the trivial cases $P(Y = -1) = 0$ or $P(Y = 1) = 0$.

Let $f_{PA}(x, \zeta_N, \hat{P}_N)$ be a function defining a *prediction algorithm* i.e. a function with values in $\{-1, 1\}$ constructed by means of $x \in \mathbb{X}$, the sample ζ_N and \hat{P}_N . In fact we consider a family of functions when instead of ζ_N we use its subsamples. Thus we write $f_{PA}(x, \zeta_N(S), \hat{P}_N)$ for $\zeta_N(S) := \{(X^j, Y^j), j \in S\}$, $S \subset (\{j_1^1, \dots, j_1^{N_1}\} \cup \{j_{-1}^1, \dots, j_{-1}^{N_{-1}}\})$. For any $f: \mathbb{X} \rightarrow \{-1, 1\}$ we try to find its estimate f_{PA} which is close in a sense to f and we employ f_{PA} to estimate $Err(f)$. For this purpose we also apply *K-fold cross-validation* procedure or, more precisely, subsampling approach. At first, for some fixed $K \in \mathbb{N}$ and each $y \in \{-1, 1\}$, we consider a partition of the set $\{j_y^1, \dots, j_y^{N_y}\}$ into K subsets $S_k^y(N_y, \omega)$, $k = 1, \dots, K$, as follows

$$S_k^y(N_y, \omega) := \left\{ j_y^i(\omega) : i \in \left\{ (k-1) \left\lfloor \frac{N_y}{K} \right\rfloor + 1, \dots, k \left\lfloor \frac{N_y}{K} \right\rfloor \right\} \cup \{k \left\lfloor \frac{N_y}{K} \right\rfloor + 1, \dots, N_y\} \right\}. \quad (9)$$

Finally, we construct $S_k(N, \omega) = S_k^1(N_1, \omega) \cup S_k^{-1}(N_{-1}, \omega)$ and introduce

$$\widehat{Err}_K(f_{PA}, \zeta_N, \hat{P}_N) := \frac{2}{K} \sum_{y \in \{-1, 1\}} \sum_{k=1}^K \sum_{j \in S_k^y(N_y)} \frac{\hat{\psi}(y, \zeta_N(\overline{S_k(N)}), \hat{P}_N) \mathbb{I}\{f_{PA}^j(N, k) \neq y\} \hat{P}_N^y}{\sharp S_k^y(N_y)} \quad (10)$$

where $f_{PA}^j(N, k) := f_{PA}(X^j, \zeta_N(\overline{S_k(N)}), \hat{P}_N)$ and $\hat{\psi}(y, \zeta_N(\overline{S_k(N)}), \hat{P}_N)$ is an estimate of $\psi(y)$, $y \in \{-1, 1\}$, constructed by observations $\zeta_N(\overline{S_k(N)})$ and \hat{P}_N ; $\overline{S_k(N)} = \{j_1^1, \dots, j_1^{N_1}\} \cup \{j_{-1}^1, \dots, j_{-1}^{N_{-1}}\} \setminus S_k(N)$ and \sharp stands for the cardinality of a finite set.

We will suppose that, for each $k = 1, \dots, K$,

$$\hat{\psi}(y, \zeta_N(\overline{S_k(N)}), \hat{P}_N) \rightarrow \psi(y) \text{ a.s., } N \rightarrow \infty, \quad y \in \{-1, 1\}. \quad (11)$$

Introduce $L(x) = \psi(1)P(X = x, Y = 1) - \psi(-1)P(X = x, Y = -1)$, $x \in \mathbb{X}$. The following result is an analogue of Theorem 1 [4] for stratified samples.

Theorem 1 *Let ζ_N be a sample described above, ψ be a penalty function, $f: \mathbb{X} \rightarrow \{-1, 1\}$ be an arbitrary function and f_{PA} define a prediction algorithm. Assume that (11) is valid and there exists a non-empty set $U \subset \mathbb{X}$ such that, for each $x \in U$ and $k = 1, \dots, K$, relation*

$$f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) \rightarrow f(x) \text{ a.s., } N \rightarrow \infty, \quad (12)$$

holds. Then, for each $a \in (0, 1)$ (with $N_1 = \max\{[aN], 1\}$, $N_{-1} = N - N_1$),

$$\widehat{Err}_K(f_{PA}, \zeta_N, \widehat{\mathbf{P}}_N) \rightarrow Err(f) \text{ a.s., } N \rightarrow \infty, \quad (13)$$

if and only if

$$\sum_{k=1}^K \sum_{y \in \{-1, 1\}} \sum_{x \in \mathbb{X}_y} y \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \widehat{\mathbf{P}}_N) = -y\} L(x) \rightarrow 0 \text{ a.s., } N \rightarrow \infty, \quad (14)$$

where

$$\mathbb{X}_y = (\mathbb{X} \setminus U) \cap \{x \in \mathbb{X} : f(x) = y\}, \quad y \in \{-1, 1\}. \quad (15)$$

Proof. It suffices to consider relation

$$\frac{2}{K} \sum_{k=1}^K \sum_{y \in \{-1, 1\}} \psi(y) \sum_{j \in S_k^y(N_y)} \frac{\mathbb{I}\{f_{PA}^j(N, k) \neq y\} \mathbf{P}(Y = y)}{\#S_k^y(N_y)} \rightarrow Err(f) \text{ a.s., } N \rightarrow \infty, \quad (16)$$

since the difference between $\widehat{Err}_K(f_{PA}, \zeta_N, \widehat{\mathbf{P}}_N)$ and the left-hand side of (16) tends to zero a.s., $N \rightarrow \infty$. Namely, we take into account (8), (11) and the inequality

$$\frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} \mathbb{I}\{f_{PA}^j(N, k) \neq y\} \leq 1$$

which is valid for each $y \in \{-1, 1\}$, any $k = 1, \dots, K$ and all N .

Now we note that in view of (1) relation (13) is equivalent to the following one

$$\frac{2}{K} \sum_{k=1}^K \sum_{y \in \{-1, 1\}} \psi(y) \left(\sum_{j \in S_k^y(N_y)} \frac{\mathbb{I}\{f_{PA}^j(N, k) \neq y\} \mathbf{P}(Y = y)}{\#S_k^y(N_y)} - \mathbf{P}(f(X) \neq y, Y = y) \right) \rightarrow 0 \text{ a.s.}$$

when $N \rightarrow \infty$.

For any $y \in \{-1, 1\}$ and $m \in \mathbb{N}$, set $\tilde{X}^{j_y^m} = \mathbb{I}\{f(X^{j_y^m}) \neq y\} - \mathbf{P}(f(X^{j_y^m}) \neq y)$. Then $|\tilde{X}^{j_y^m}| \leq 2$, $\mathbf{E}\tilde{X}^{j_y^m} = 0$ for such y and m . Fix any $y \in \{-1, 1\}$ and $k \in \{1, \dots, K\}$. Consider an array $\mathcal{A}(y, k) := \{\tilde{X}^{j_y^m}, j_y^m \in S_k^y(N_y), N_y = N_y(N)\}$ where $N \in \mathbb{N}$. The strong law of large numbers for arrays (SLLNA) given in Theorem 2.1 [21] applies, e.g., when $p = 2$, $\psi(x) = |x|^{5/2}$, $x \in \mathbb{R}$, and $k \in \mathbb{N}$ (we keep here the notation p , ψ and k used in the mentioned paper [21] for other objects). Whence one has

$$\frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} \mathbb{I}\{f(X^j) \neq y\} \rightarrow \mathbf{P}(f(X^{j_y^1}) \neq y) \text{ a.s., } N \rightarrow \infty. \quad (17)$$

More precisely, in [21] a triangular array was considered. The study of our array $\mathcal{A}(y, k)$ can be reduced to analysis of a collection of several triangular arrays. Indeed, in view of (9) the contribution to asymptotic behavior of $\frac{1}{\#S_K^y(N_y)} \sum_{j \in S_K^y(N_y)} \mathbb{I}\{f(X^j) \neq y\}$ of the last summands appearing in $S_K^y(N_y)$ with indices greater than $[N_y/K]K$ is negligible. Therefore, for each k belonging to $\{1, \dots, K\}$ we have an array $\tilde{\mathcal{A}}(y, k)$ with rows containing r_1, r_2, \dots elements $((r_i)_{i \in \mathbb{N}}$ depends on $(N_y(N))_{N \in \mathbb{N}}$) such that $r_1 \leq r_2 \leq \dots$ ($\tilde{\mathcal{A}}(y, k) = \mathcal{A}(y, k)$ for

$k = 1, \dots, K - 1$). Moreover, if $r_m = \dots = r_q$ then $q - m \leq I$ where $I = [Ka]$. Now we can consider separately each of I arrays (taking one row among the I rows of equal length) with strictly increasing numbers of elements in rows. Then we introduce auxiliary rows, if necessary, with elements $\mathbb{P}(f(X^{j_y^1}) \neq y)$ to get a triangular array and apply the mentioned SLLNA. Lemma 1 shows that $\mathbb{P}(f(X^{j_y^1}) \neq y) = \mathbb{P}(f(X) \neq y | Y = y)$. Hence

$$\frac{2}{K} \sum_{k=1}^K \sum_{y \in \{-1, 1\}} \psi(y) \sum_{j \in S_k^y(N_y)} \frac{\mathbb{I}\{f(X^j) \neq y\} \mathbb{P}(Y = y)}{\#S_k^y(N_y)} \rightarrow Err(f) \text{ a.s., } N \rightarrow \infty.$$

Therefore, relation (16) is valid if and only if

$$\sum_{k=1}^K \sum_{y \in \{-1, 1\}} \psi(y) \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} Z_{N,k}^j(y) \rightarrow 0 \text{ a.s., } N \rightarrow \infty,$$

where

$$Z_{N,k}^j(y) = (\mathbb{I}\{f_{PA}(X^j, \zeta_N(\overline{S_k(N)}), \widehat{\mathbf{P}}_N) \neq y\} - \mathbb{I}\{f(X^j) \neq y\}) \mathbb{P}(Y = y), \quad y \in \{-1, 1\}.$$

Set $F_{N,k}(x, y) := \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \widehat{\mathbf{P}}_N) \neq y\} - \mathbb{I}\{f(x) \neq y\}$. For each $y \in \{-1, 1\}$, $N \in \mathbb{N}$ and $k = 1, \dots, K$, we introduce the random variables

$$Q_{N,k,y} = \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} F_{N,k}(X^j, y).$$

Then (16) is equivalent to the following relation

$$\sum_{k=1}^K \sum_{y \in \{-1, 1\}} \psi(y) \mathbb{P}(Y = y) Q_{N,k,y} \rightarrow 0 \text{ a.s., } N \rightarrow \infty.$$

Note that $Q_{N,k,y} = Q_{N,k,y}^{(1)} + Q_{N,k,y}^{(2)}$ where

$$Q_{N,k,y}^{(1)} = \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} \mathbb{I}\{X^j \in U\} F_{N,k}(X^j, y),$$

$$Q_{N,k,y}^{(2)} = \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} \mathbb{I}\{X^j \notin U\} F_{N,k}(X^j, y).$$

We have

$$|Q_{N,k,y}^{(1)}| \leq \sum_{x \in U} \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} |\mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \widehat{\mathbf{P}}_N) \neq y\} - \mathbb{I}\{f(x) \neq y\}|.$$

Condition (12) entails

$$\sum_{k=1}^K \sum_{y \in \{-1, 1\}} \psi(y) \mathbb{P}(Y = y) Q_{N,k,y}^{(1)} \rightarrow 0 \text{ a.s., } N \rightarrow \infty.$$

If $U = \mathbb{X}$, then $Q_{N,k,y}^{(2)} = 0$ for all N, k, y and (16) holds. Let $U \neq \mathbb{X}$. Then

$$V_{N,k} := \sum_{y \in \{-1,1\}} \psi(y) \mathbb{P}(Y = y) Q_{N,k,y}^{(2)} = \sum_{x \in \mathbb{X}_{-1} \cup \mathbb{X}_1} \sum_{y \in \{-1,1\}} G_{N,k,y}(x)$$

where $N \in \mathbb{N}$, $k = 1, \dots, K$, the sets \mathbb{X}_y were introduced in (15) and

$$G_{N,k,y}(x) := \frac{\psi(y) \mathbb{P}(Y = y)}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} \mathbb{I}\{X^j = x\} \left(\mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{\mathbf{P}}_N) \neq y\} - \mathbb{I}\{f(x) \neq y\} \right).$$

Let $R_j^y(x) := \mathbb{I}\{X^j = x\} \psi(y) \mathbb{P}(Y = y)$, $y \in \{-1, 1\}$, $j \in \mathbb{N}$. The following equalities are valid

$$\sum_{x \in \mathbb{X}_1} \sum_{y \in \{-1,1\}} G_{N,k,y}(x) = \sum_{x \in \mathbb{X}_1} \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{\mathbf{P}}_N) = -1\} \sum_{y \in \{-1,1\}} \frac{y}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} R_j^y(x),$$

$$\sum_{x \in \mathbb{X}_{-1}} \sum_{y \in \{-1,1\}} G_{N,k,y}(x) = - \sum_{x \in \mathbb{X}_{-1}} \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{\mathbf{P}}_N) = 1\} \sum_{y \in \{-1,1\}} \frac{y}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} R_j^y(x).$$

Similarly to (17) it can be shown that, for each $y \in \{-1, 1\}$, $k = 1, \dots, K$ and $x \in \mathbb{X}$, relation

$$\frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} R_j^y(x) \rightarrow J(x, y) \text{ a.s., } N \rightarrow \infty,$$

holds where $J(x, y) = \psi(y) \mathbb{P}(X = x, Y = y)$.

Thus, for any $\varepsilon > 0$, $x \in \mathbb{X}$, $y \in \{-1, 1\}$ and almost all $\omega \in \Omega$, there exists $N_1(\omega, \varepsilon, k, y)$ such that

$$\left| \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} (R_j^y(x) - J(x, y)) \right| < \varepsilon \quad (18)$$

for $N > N_1(\omega, \varepsilon, k, y)$. We note that $L(x) = \sum_{y \in \{-1,1\}} y J(x, y)$. Furthermore,

$$\begin{aligned} V_{N,k} &= \sum_{x \in \mathbb{X}_1} \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{\mathbf{P}}_N) = -1\} L(x) \\ &+ \sum_{x \in \mathbb{X}_1} \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{\mathbf{P}}_N) = -1\} \sum_{y \in \{-1,1\}} \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} y (R_j^y(x) - J(x, y)) \\ &- \sum_{x \in \mathbb{X}_{-1}} \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{\mathbf{P}}_N) = 1\} L(x) \\ &- \sum_{x \in \mathbb{X}_{-1}} \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{\mathbf{P}}_N) = 1\} \sum_{y \in \{-1,1\}} \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} y (R_j^y(x) - J(x, y)). \end{aligned}$$

Taking into account (18), for any $\varepsilon > 0$, we obtain the inequality

$$\left| \sum_{k=1}^K \left(\sum_{x \in \mathbb{X}_1} \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) = -1\} \sum_{y \in \{-1, 1\}} \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} y(R_j^y(x) - J(x, y)) \right. \right. \\ \left. \left. - \sum_{x \in \mathbb{X}_{-1}} \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) = 1\} \sum_{y \in \{-1, 1\}} \frac{1}{\#S_k^y(N_y)} \sum_{j \in S_k^y(N_y)} y(R_j^y(x) - J(x, y)) \right) \right| < 2K\varepsilon\#\mathbb{X}$$

when $N > N_1(\omega, \varepsilon) := \max_{k=1, \dots, K, y \in \{-1, 1\}} N_1(\omega, \varepsilon, k, y)$. Thus, $\sum_{k=1}^K V_{N,k} \rightarrow 0$ a.s., as $N \rightarrow \infty$, if and only if condition (14) holds. \square

Remark 2. Theorem 1 demonstrates what one has to verify outside the “good set” U (where f_{PA} approximates f point-wise) to guarantee validity of (13). To clarify (14) we can rewrite it (cf. Theorem 1 [9]) in the following way

$$\sum_{k=1}^K \sum_{y \in \{-1, 1\}} \sum_{x \in \mathbb{X}_y} y \mathbb{I}\{f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) \neq f(x)\} L(x) \rightarrow 0 \text{ a.s., } N \rightarrow \infty.$$

Note also that condition (14) coincides with the corresponding one proposed for the case of i.i.d. observations without stratification (cf. Theorem 1 [4]).

Remark 3. According to Corollary 1 [4] the choice of

$$U = \left\{ x \in M : P(Y = 1|X = x) \neq \frac{\psi(-1)}{\psi(-1) + \psi(1)} \right\}, \quad (19)$$

where M appears in (2), implies that (14) is satisfied and the problem is to prove that (12) holds on this U for f under consideration and appropriate f_{PA} .

Remark 4. There are various ways to get strongly consistent estimates \hat{P}_N^y , $y \in \{-1, 1\}$. The first one is to estimate $P(Y = y)$, $y \in \{-1, 1\}$, by means of another sample. In this case it is essential to have the samples belonging to the same law. The second approach involves estimates of $P(Y = y)$ constructed simultaneously with forming of the sample ζ_N . More exactly, \hat{P}_N^y is a frequency estimate of the $P(Y = y)$ by means of a random number of observations $Y^1, Y^2, \dots, Y^{\tilde{N}}$ where $\tilde{N} = \max\{j_1^{N_1}, j_{-1}^{N_{-1}}\}$. Such estimate is strongly consistent as $N \rightarrow \infty$ since $\tilde{N} \geq N$ a.s. We have mentioned in Section 1 that the choice of ψ given by (3) is natural. Moreover, for such ψ we can simplify (10). Namely, now it need not contain $\hat{\psi}(y, \zeta_N(\overline{S_k(N)}), \hat{P}_N)$ and \hat{P}_N^y .

The estimates of $P(X \in B, Y = y)$ where $B \subset \mathbb{X}$ and $y \in \{-1, 1\}$ are required if $\psi(y)$ depends on the joint distribution of X and Y . For this estimation we cannot employ the sample ζ_N . However the Bayes formula can help to construct the desired estimates. Such approach in the framework of classification problems was employed by J.Park in [17].

Example. We show how for the stratified sample ζ_N and an optimal function f^* one can find f_{PA} to employ the result described by Theorem 1.

For $B \subset \mathbb{X}$ the Bayes theorem yields

$$P(Y = 1|X \in B) = \frac{P(X \in B|Y = 1)P(Y = 1)}{P(X \in B|Y = 1)P(Y = 1) + P(X \in B|Y = -1)P(Y = -1)}.$$

Therefore, one can construct the estimates of $P(Y = 1|X \in B)$ by means of ζ_N and \hat{P}_N . Plug-in principle suggests us how to get the appropriate f_{PA} for f^* . According to this principle we construct estimates of $P(Y = 1|X = x)$ and $\frac{\psi(-1)}{\psi(-1)+\psi(1)}$. Then we substitute them into the definition of A introduced by (2).

For $y \in \{-1, 1\}$, $k = 1, \dots, K$ and $N \in \mathbb{N}$, set $W_k^y(N_y) := \cup_{m \neq k} S_m^y(N_y)$ and assume that (11) holds. Consider the following estimates for $P(Y = 1|X = x)$ and $\frac{\psi(-1)}{\psi(-1)+\psi(1)}$, respectively

$$\begin{aligned} g(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) &:= \frac{\hat{P}_N^1 I_1(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N)}{\hat{P}_N^{-1} I_{-1}(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) + \hat{P}_N^1 I_1(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N)}, \\ h(\zeta_N(\overline{S_k(N)}), \hat{P}_N) &:= \frac{\hat{\psi}(-1, \zeta_N(\overline{S_k(N)}), \hat{P}_N)}{\hat{\psi}(-1, \zeta_N(\overline{S_k(N)}), \hat{P}_N) + \hat{\psi}(1, \zeta_N(\overline{S_k(N)}), \hat{P}_N)} \end{aligned} \quad (20)$$

where

$$I_y(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) := \frac{1}{\#W_k^y(N_y)} \sum_{j \in W_k^y(N_y)} \mathbb{I}(X^j = x), \quad y \in \{-1, 1\}, \quad x \in \mathbb{X}.$$

As usual we set formally $0/0 := 0$. Let us define

$$A_N = \{x \in \mathbb{X}: g(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) > h(\zeta_N(\overline{S_k(N)}), \hat{P}_N)\}$$

and

$$f_{PA}(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) := \mathbb{I}\{A_N\}(x) - \mathbb{I}\{\overline{A_N}\}(x). \quad (21)$$

It can be established similarly to the proof of (17) that for $y \in \{-1, 1\}$ and $x \in \mathbb{X}$

$$I_y(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) \rightarrow P(X = x|Y = y) \text{ a.s., } N \rightarrow \infty.$$

By virtue of (8) and (11) we come to the following relations

$$\begin{aligned} g(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) &\rightarrow P(Y = 1|X = x) \text{ a.s., } N \rightarrow \infty, \\ h(\zeta_N(\overline{S_k(N)}), \hat{P}_N) &\rightarrow \frac{\psi(-1)}{\psi(-1) + \psi(1)} \text{ a.s., } N \rightarrow \infty. \end{aligned}$$

For any $x \in A$ one has $P(Y = 1|X = x) > \frac{\psi(-1)}{\psi(-1)+\psi(1)}$. Therefore, for every $\varepsilon > 0$ such that

$$P(Y = 1|X = x) - \varepsilon > \frac{\psi(-1)}{\psi(-1) + \psi(1)} + \varepsilon,$$

for almost all $\omega \in \Omega$ and any $x \in A$ there exist $N_2(\omega)$ and $N_3(\omega)$ such that

$$g(x, \zeta_N(\overline{S_k(N)}), \hat{P}_N) > P(Y = 1|X = x) - \varepsilon$$

for each $N > N_2(\omega)$ and

$$h(\zeta_N(\overline{S_k(N)}), \hat{P}_N) < \frac{\psi(-1)}{\psi(-1) + \psi(1)} + \varepsilon$$

for each $N > N_3(\omega)$. Thus, for almost all $\omega \in \Omega$, any $x \in A$ and $N > \max\{N_2(\omega), N_3(\omega)\}$ the equalities $f(x) = 1$ and $f_{PA}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N) = 1$ are true. Similarly, for

$$x \in U \cap \overline{A} = \left\{ x \in M : P(Y = 1|X = x) < \frac{\psi(-1)}{\psi(-1) + \psi(1)} \right\}$$

it can be shown that for almost all $\omega \in \Omega$ there exists such $N_4(\omega) \in \mathbb{N}$ that, for $N > N_4(\omega)$, the equalities $f(x) = -1$ and $f_{PA}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N) = -1$ are satisfied. Thus condition (12) holds for U introduced by formula (19). Hence, the estimate of $Err(f^*)$ appearing in Theorem 1 is strongly consistent if $f_{PA}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N)$ is defined by (21). The desired f_{PA} is constructed. \square

If ψ is introduced by (3) then U and A have the following form

$$\begin{aligned} U &= \{x \in M : P(Y = 1|X = x) \neq P(Y = 1)\}, \\ A &= \{x \in M : P(Y = 1|X = x) > P(Y = 1)\}. \end{aligned}$$

Consequently, instead of (20) we can set $h(\zeta_N(\overline{S_k(N)}), \widehat{P}_N) := \widehat{P}_N^1$ and simplify our Example.

Now we turn to the situation when a response variable Y depends only on the part of factors. Let $\{k_1, \dots, k_r\}$ be a subset of $\{1, \dots, n\}$ where $r < n$. We say (cf. [2], Ch. 2) that the factors X_{k_1}, \dots, X_{k_r} are relevant (or collection of indices k_1, \dots, k_r is relevant) whenever for each $x = (x_1, \dots, x_n) \in M$ it turns out that

$$P(Y = 1|X_1 = x_1, \dots, X_n = x_n) = P(Y = 1|X_{k_1} = x_{k_1}, \dots, X_{k_r} = x_{k_r}). \quad (22)$$

For an arbitrary collection $\{m_1, \dots, m_r\} \subset \{1, \dots, n\}$ set

$$f^{m_1, \dots, m_r}(x) = \begin{cases} 1, & P(Y = 1|X_{m_1} = x_{m_1}, \dots, X_{m_r} = x_{m_r}) > \frac{\psi(-1)}{\psi(-1) + \psi(1)}, \quad x \in M, \\ -1, & \text{otherwise.} \end{cases}$$

If $\{k_1, \dots, k_r\}$ is a relevant collection then the optimal function f^* has the form f^{k_1, \dots, k_r} . Therefore, for any relevant collection $\{k_1, \dots, k_r\} \subset \{1, \dots, n\}$ and an arbitrary subset $\{m_1, \dots, m_r\} \subset \{1, \dots, n\}$, the following inequality holds

$$Err(f^{k_1, \dots, k_r}) \leq Err(f^{m_1, \dots, m_r}).$$

For any $\{m_1, \dots, m_r\} \subset \{1, \dots, n\}$, $x \in \mathbb{X}$ and a penalty function ψ , consider a prediction algorithm with $f_{PA}^{m_1, \dots, m_r}$ such that

$$f_{PA}^{m_1, \dots, m_r}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N) = \begin{cases} 1, & g^{m_1, \dots, m_r}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N) > h(\zeta_N(\overline{S_k(N)}), \widehat{P}_N), \\ -1, & \text{otherwise,} \end{cases}$$

where $k = 1, \dots, K$, h is defined by (20) and

$$g^{m_1, \dots, m_r}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N) = \frac{\widehat{P}_N^1 I_1^{m_1, \dots, m_r}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N)}{\widehat{P}_N^{-1} I_{-1}^{m_1, \dots, m_r}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N) + \widehat{P}_N^1 I_1^{m_1, \dots, m_r}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N)}.$$

Here, for $y \in \{-1, 1\}$ and $x \in \mathbb{X}$,

$$I_y^{m_1, \dots, m_r}(x, \zeta_N(\overline{S_k(N)}), \widehat{P}_N) = \frac{1}{\#W_k^y(N_y)} \sum_{j \in W_k^y(N_y)} \mathbb{I}(X_{m_1}^j = x_{m_1}, \dots, X_{m_r}^j = x_{m_r}).$$

Set

$$U := \left\{ x \in M_{m_1, \dots, m_r} : \mathbf{P}(Y = 1 | X_{m_1} = x_{m_1}, \dots, X_{m_r} = x_{m_r}) \neq \frac{\psi(-1)}{\psi(-1) + \psi(1)} \right\}$$

where $M_{m_1, \dots, m_r} = \{x \in \mathbb{X} : \mathbf{P}(X_{m_1} = x_{m_1}, \dots, X_{m_r} = x_{m_r}) > 0\}$. Similarly to [4], it can be shown that for relevant collections $\{k_1, \dots, k_r\} \subset \{1, \dots, n\}$, arbitrary collections $\{m_1, \dots, m_r\} \subset \{1, \dots, n\}$, any $\varepsilon > 0$ and almost all $\omega \in \Omega$,

$$\widehat{Err}_K(f_{PA}^{k_1, \dots, k_r}, \zeta_N, \widehat{\mathbf{P}}_N) \leq \widehat{Err}_K(f_{PA}^{m_1, \dots, m_r}, \zeta_N, \widehat{\mathbf{P}}_N) + \varepsilon \text{ a.s.}$$

when N is large enough. Thus, it is natural to choose as relevant collection of factors X_{k_1}, \dots, X_{k_r} that one which has the minimal prediction error estimate $\widehat{Err}_K(f_{PA}^{k_1, \dots, k_r}, \zeta_N, \widehat{\mathbf{P}}_N)$.

4 Cost approach to experiments and the XOR-model

We compare two approaches concerning the application of the MDR-EFE method for different sample plans. The first one was described in [4] and consists in the employment of nonrandom number of i.i.d. observations. The second one is considered in this paper and involves the stratified sample. More exactly, stratification means the separation of observations taking into account the values of a response variable under consideration. We will denote these methods as iMDR-EFE and sMDR-EFE, respectively. It seems that the main disadvantage of the second approach is that too large number of independent observations is required to form a stratified sample of N elements with fixed cases to controls ratio when $\mathbf{P}(Y = 1)$ is very small. Moreover, we have to skip a lot of observations with $Y^i = -1$. However it is worth to emphasize that no information on predictors X^i is needed at the stage of forming a stratified sample, therefore we need not get X^i for skipped observations. This is essential when the measurement of Y is cheaper than the measurement of X (since X has components X_1, \dots, X_n). Thus it is not interesting to compare iMDR-EFE and sMDR-EFE for the equal sizes of samples and we should take into account additional observations in sMDR-EFE.

We propose to compare two approaches in the sense of the total cost of experiment. Assume that there is a fixed amount of money C ($C \in \mathbb{N}$) for research. Let each observation (X^i, Y^i) cost 1 and let the ratio of the price of measuring Y to that of X be $w \in \mathbb{R}_+$. We consider the maximal sizes $s_{ind}(C, w)$ and $s_{str}(C, w)$ of the samples which are available in experiments organized to apply iMDR-EFE and sMDR-EFE, respectively. Let $C_{ind}(N, w)$ be the total cost of the sample having size N and consisting of independent observations. Let $C_{str}(N, w)$ be the total cost of the stratified sample of size N . Then

$$\begin{aligned} s_{ind}(C, w) &= \max\{N \in \mathbb{N} : C_{ind}(N, w) \leq C\}, \\ s_{str}(C, w) &= \max\{N \in \mathbb{N} : C_{str}(N, w) \leq C\}. \end{aligned}$$

Clearly, $C_{ind}(N, w) = N$ and therefore $s_{ind}(C, w) = C$. We note that this value is nonrandom and is known before the experiment.

Recall (see Remark 4) that $\tilde{N} = \tilde{N}(N, a, \omega) = \max\{j_{-1}^{N-1}, j_1^{N_1}\}$. Hence

$$C_{str}(N, w, \omega) = \frac{1}{w+1}N + \frac{w}{w+1}\tilde{N}(N, a, \omega)$$

because we have to measure X in N observations (which are included into the sample) and to measure Y in \tilde{N} observations. Thus $C_{str}(N, w)$ is a random variable which is unknown before the experiment. In general case, $s_{str}(C, w)$ is a random variable as well. However we can select such N that $C_{str}(N, w)$ does not exceed C with probability which is not less than $1 - \alpha$, $\alpha \in (0, 1)$. We assume that if the cost of experiment exceeds C with given small probability then some Reserve Foundation will cover the extra expenses. Now we want to find $s'_{str}(C, w, \alpha)$ so that

$$s'_{str}(C, w, \alpha) = \max \{N \in \mathbb{N}: \mathbf{P}(C_{str}(N, w) \leq C) \geq 1 - \alpha\}.$$

Obviously, $\tilde{N}(N, a, \omega) \geq N$ for each $\omega \in \Omega$. Therefore for any $N \in \mathbb{N}$ and $\omega \in \Omega$

$$C_{str}(N, w, \omega) = N + \frac{w}{w+1}(\tilde{N}(N, a, \omega) - N) \geq N.$$

Consequently, for any $N > C$,

$$\mathbf{P}(C_{str}(N, w) \leq C) \leq \mathbf{P}(C_{str}(N, w) < N) = 0.$$

So, for arbitrary $C \in \mathbb{N}$, $\alpha \in (0, 1)$ and $w > 0$, we observe that

$$s'_{str}(C, w, \alpha) = \max \{N \in \{1, 2, \dots, C\}: \mathbf{P}(C_{str}(N, w) \leq C) \geq 1 - \alpha\}. \quad (23)$$

It is worth mentioning that $s'_{str}(C, w, \alpha)$ is a non-random variable which depends on certain parameters and the distribution of \tilde{N} . We will apply the following result.

Lemma 2 *For each $m \in \{0, 1, 2, \dots\}$ one has*

$$\mathbf{P}(\tilde{N} = m) = \begin{cases} 0, & m < N, \\ \mathbf{P}(\eta_{-1} = m - N_{-1}) + \mathbf{P}(\eta_1 = m - N_1), & \text{otherwise} \end{cases} \quad (24)$$

where $\eta_{-1} \sim NB(N_{-1}, \mathbf{P}(Y = 1))$, $\eta_1 \sim NB(N_1, \mathbf{P}(Y = -1))$.

Proof. $\mathbf{P}(\tilde{N} = m) = 0$ for each $m \in \{0, 1, 2, \dots, N - 1\}$ since $\tilde{N}(N, a, \omega) \geq N$ for each $\omega \in \Omega$. If $m \in \{N, N + 1, \dots\}$ then

$$\begin{aligned} \mathbf{P}(\tilde{N} = m) &= \mathbf{P}(\max\{j_{-1}^{N-1}, j_1^{N_1}\} = m) \\ &= \mathbf{P}(j_{-1}^{N-1} = m, j_1^{N_1} < m) + \mathbf{P}(j_{-1}^{N-1} < m, j_1^{N_1} = m) \\ &\quad + \mathbf{P}(j_{-1}^{N-1} = m, j_1^{N_1} = m). \end{aligned}$$

Let $\tilde{S}_y(k) = \#\{i \in \{1, \dots, k\}: Y^i = y\}$. We note that $\tilde{S}_y(k) + \tilde{S}_{-y}(k) = k$ for each $k \in \mathbb{N}$. Thus, for each $m \in \{N, N + 1, \dots\}$ and any $y \in \{-1, 1\}$,

$$\begin{aligned} \{j_y^{N_y} = m\} &= \{\tilde{S}_y(m) = N_y\} \cap \{\tilde{S}_y(m-1) = N_y - 1\} \subset \{\tilde{S}_y(m-1) = N_y - 1\} \\ &= \{\tilde{S}_{-y}(m-1) = m - N_y\} \subset \{\tilde{S}_{-y}(m-1) \geq N_{-y}\} \subset \{j_{-y}^{N_{-y}} < m\}. \end{aligned}$$

Moreover, $j_{-1}^{N-1}(\omega) \neq j_1^{N_1}(\omega)$ for any $\omega \in \Omega$. Therefore

$$\mathbf{P}(\tilde{N} = m) = \mathbf{P}(j_{-1}^{N-1} = m) + \mathbf{P}(j_1^{N_1} = m).$$

According to Section 2, $j_y^{N_y}$ has the same distribution as $N_y + \eta_y$. Thus (24) holds. \square

Remark 5. Lemma 2 shows that the law of \tilde{N} depends on the known parameters N_{-1} , N_1 and, in general, unknown $\mathbf{P}(Y = 1)$. However, if $\mathbf{P}(Y = 1)$ is known or we have its estimates constructed by means either of $\xi_{\tilde{N}}$ or another sample (see Remark 4) then $s'_{str}(C, w, \alpha)$ can be evaluated or estimated.

Using Lemma 2 we can rewrite $\mathbf{P}(C_{str}(N, w) \leq C)$ as

$$\begin{aligned} \mathbf{P}(C_{str}(N, w) \leq C) &= \mathbf{P}\left(\tilde{N}(N, a, \omega) \leq \frac{w+1}{w}C - \frac{N}{w}\right) \\ &= \mathbf{P}\left(N_1 \leq \eta_{-1} \leq \frac{w+1}{w}C - \frac{N}{w} - N_{-1}\right) + \mathbf{P}\left(N_{-1} \leq \eta_1 \leq \frac{w+1}{w}C - \frac{N}{w} - N_1\right) \end{aligned} \quad (25)$$

where η_{-1} and η_1 are defined above. Thus, for any N , the probability $\mathbf{P}(C_{str}(N, w) \leq C)$ can be evaluated or estimated (see Remark 5). Moreover,

$$\mathbf{P}(C_{str}(N, w) \leq C) \geq \mathbf{P}(C_{str}(N+1, w) \leq C) \quad (26)$$

for all $N \in \mathbb{N}$ since, for each $\omega \in \Omega$,

$$\frac{1}{w+1}N + \frac{w}{w+1}\tilde{N}(N, a, \omega) \leq \frac{1}{w+1}(N+1) + \frac{w}{w+1}\tilde{N}(N+1, a, \omega).$$

Formulae (25) and (26) suggest the following algorithm. If $p := \mathbf{P}(Y = 1)$ is known we evaluate $\mathbf{P}(C_{str}(N, w) \leq C)$ by (25) for each $N = 1, 2, \dots$ (till C) and identify the maximal N belonging to $\{1, \dots, C\}$ such that (23) holds. In this way we determine $s'_{str}(C, w, \alpha)$. If p is unknown but there exists its estimate $\hat{p}_{N'}$ such that $\hat{p}_{N'} \rightarrow p$ almost surely as $N' \rightarrow \infty$ (for instance, it may be an estimate by means of another sample) then for each fixed C , N and w we can estimate $\mathbf{P}(C_{str}(N, w) \leq C)$ by

$$\nu_{N'}(N, w, C) = \sum_{N_1 \leq m \leq L_1} \binom{N_1 + m - 1}{m} (1 - \hat{p}_{N'})^{N-1} \hat{p}_{N'}^m + \sum_{N_{-1} \leq m \leq L_{-1}} \binom{N_{-1} + m - 1}{m} \hat{p}_{N'}^{N_1} (1 - \hat{p}_{N'})^m$$

where $L_y := \frac{w+1}{w}C - \frac{N}{w} - N_{-y}$, $y \in \{-1, 1\}$. Indeed, $\nu_{N'}(N, w, C)$ tends to the right-hand side of (25) almost surely as $N' \rightarrow \infty$. Thus we can introduce

$$\hat{s}'_{str, N'}(C, w, \alpha) := \max \{N \in \{1, 2, \dots, C\} : \nu_{N'}(N, w, C) \geq 1 - \alpha\}$$

and apply the approach proposed for known p .

To compare iMDR-EFE and sMDR-EFE we turn to the popular XOR-model (see, e.g., [23]). This model is used in genetics to describe epistasis without main effects. Namely, let $\mathbb{X} = \{0, 1, 2\}^n$ ($0, 1, 2$ correspond to the number of minor alleles of a specified gene). Assume now that the components of a random vector $X = (X_1, \dots, X_n)$ are independent and, for each $i \in \{1, \dots, n\}$, there exists such $p_i \in (0, 0.5]$ that

$$\mathbf{P}(X_i = 0) = (1 - p_i)^2, \quad \mathbf{P}(X_i = 1) = 2p_i(1 - p_i), \quad \mathbf{P}(X_i = 2) = p_i^2. \quad (27)$$

This situation is typical for genome-wide association studies (GWAS) where each X_i corresponds to a single nucleotide polymorphism (SNP) and p_i is a minor allele frequency (MAF). Suppose that collection of relevant factors (describing a binary response variable Y) is X_{k_1}, \dots, X_{k_r} where $\{k_1, \dots, k_r\} \subset \{1, \dots, n\}$. One also says that collection of indices $\bar{k} := \{k_1, \dots, k_r\}$ is relevant.

In our simulations we employ the XOR-model of dependence between predictors and response variable. It is a generalization of the XOR-model described in [23] to the case of more than 2 relevant factors. Namely, for each $x = (x_1, \dots, x_n) \in \mathbb{X}$,

$$P(Y = 1|X = x) = \begin{cases} \gamma, & (x_{k_1} + \dots + x_{k_r}) \bmod 2 = 1, \\ 0, & \text{otherwise} \end{cases} \quad (28)$$

where $\gamma \in (0, 1)$ and $p_{k_1} = \dots = p_{k_r} = 0.5$. Thus $(X_{k_1}, \dots, X_{k_r})$ is a collection of relevant factors according to (22). XOR-model is interesting as it possess the property described by Lemma 3 below.

For any $\bar{s} := \{s_1, \dots, s_q\} \subset \{1, \dots, n\}$, $q \leq n$, set $X_{\bar{s}} := (X_{s_1}, \dots, X_{s_q})$. We write $X_{\bar{s}} = x_{\bar{s}}$ where $x_{\bar{s}} = (x_{s_1}, \dots, x_{s_q}) \in \{0, 1, 2\}^q$ if $X_{s_i} = x_{s_i}$ for all $i = 1, \dots, q$.

Lemma 3 *Let \mathbb{X}, X, Y and \bar{k} be introduced above (the dependence between X and Y is described by (28)). Then, for any collection $\bar{m} = \{m_1, \dots, m_l\} \subset \{1, \dots, n\}$, a response variable Y is dependent with $X_{\bar{m}}$ if and only if $\bar{k} \subset \bar{m}$.*

Proof. Formula (28) shows that Y and $X_{\bar{m}}$ are dependent if $\bar{k} \subset \bar{m}$. Let now $\bar{m} = \bar{u} \cup \bar{v}$ where $\bar{u} \subsetneq \bar{k}$ and $\bar{v} \subsetneq \{1, \dots, n\} \setminus \bar{k}$. Introduce $t := \#\bar{u}$. Since $P(X_{\bar{m}} = x_{\bar{m}}) \neq 0$ for each $x_{\bar{m}} \in \{0, 1, 2\}^l$ it remains to establish that

$$P(Y = 1|X_{\bar{m}} = x_{\bar{m}}) = P(Y = 1). \quad (29)$$

Evidently

$$P(Y = 1|X_{\bar{m}} = x_{\bar{m}}) \quad (30)$$

$$= \sum_{x_{\bar{k} \setminus \bar{u}} \in \{0, 1, 2\}^{r-t}} P(Y = 1|X_{\bar{u}} = x_{\bar{u}}, X_{\bar{k} \setminus \bar{u}} = x_{\bar{k} \setminus \bar{u}}, X_{\bar{v}} = x_{\bar{v}}) P(X_{\bar{k} \setminus \bar{u}} = x_{\bar{k} \setminus \bar{u}}) \quad (31)$$

$$= \gamma \sum_{x_{\bar{k} \setminus \bar{u}} \in \{0, 1, 2\}^{r-t}} \mathbb{I} \left\{ \left(\sum_{i \in \bar{k} \setminus \bar{u}} x_i + \sum_{i \in \bar{u}} x_i \right) \bmod 2 = 1 \right\} P(X_{\bar{k} \setminus \bar{u}} = x_{\bar{k} \setminus \bar{u}}) \quad (32)$$

$$= \gamma P \left(\sum_{i \in \bar{k} \setminus \bar{u}} X_i \bmod 2 \neq \sum_{i \in \bar{u}} x_i \bmod 2 \right) \quad (33)$$

In view of (27) the laws of (X_1, \dots, X_n) and (X'_1, \dots, X'_n) coincide where $X'_i = \sum_{j=1}^2 B_i^j$ and $B_1^1, B_1^2, \dots, B_n^1, B_n^2$ are independent Bernoulli variables such that $P(B_i^k = 1) = p_i$, $k = 1, 2$, $i = 1, \dots, n$. Hence, $\text{law}(\sum_{i \in \bar{s}} X_i) = \text{law}(\sum_{i \in \bar{s}} (B_i^1 + B_i^2))$ for any $\bar{s} \subset \{1, \dots, n\}$. Thus, for $\bar{s} := \{s_1, \dots, s_q\} \subset \bar{k}$ ($q \leq r$), one has

$$P \left(\sum_{i \in \bar{s}} X_i \bmod 2 = 0 \right) = P \left(\sum_{i=1}^{2q} \tilde{B}_i \bmod 2 = 0 \right) = \sum_{j=0}^q C_{2q}^{2j} 2^{-2q} = \frac{1}{2}$$

where $\tilde{B}_1, \dots, \tilde{B}_{2r}$ are i.i.d. Bernoulli random variables with probability of success 0.5. Therefore,

$$\mathbb{P} \left(\sum_{i \in \bar{s}} X_i \bmod 2 = 1 \right) = \frac{1}{2}.$$

Taking into account (33) we observe that $\mathbb{P}(Y = 1 | X_{\bar{m}} = x_{\bar{m}}) = \frac{\gamma}{2}$ for any $x_{\bar{m}} \in \{0, 1, 2\}^l$. To complete the proof note that

$$\begin{aligned} \mathbb{P}(Y = 1) &= \sum_{x \in \mathbb{X}} \mathbb{P}(Y = 1 | X = x) \mathbb{P}(X = x) \\ &= \sum_{x_{\bar{k}} \in \{0, 1, 2\}^r} \mathbb{P}(Y = 1 | X_{\bar{k}} = x_{\bar{k}}) \mathbb{P}(X_{\bar{k}} = x_{\bar{k}}) = \gamma \mathbb{P} \left(\sum_{i \in \bar{k}} X_i \bmod 2 = 1 \right) = \frac{\gamma}{2}. \end{aligned}$$

Thus (29) holds. \square

Remark 6. Lemma 3 shows that only the whole collection of relevant factors X_{k_1}, \dots, X_{k_r} and not some of its subcollections determines the response Y in XOR-model.

To complete this Section we discuss some asymptotical properties of a random sample as $C \rightarrow \infty$. For a given C the iMDR-EFE method involves C observations whereas its sMDR-EFE counterpart (using the same amount of money C) leads to N observations and

$$\frac{1}{w+1}N + \frac{w}{w+1} \max\{j_{-1}^{N - ([aN] \vee 1)}, j_1^{[aN] \vee 1}\} \leq C, \quad (34)$$

here $a \in (0, 1)$. Clearly, (34) implies that $N \leq C$. Now we consider the problem whether one can take $N = \lambda C$ for some $\lambda \in (0, 1)$ such that with probability close to one inequality (34) is satisfied.

Lemma 4 *For an arbitrary (fixed) $a \in (0, 1)$ and w inequality (34) is valid with probability tending to one as $C \rightarrow \infty$ if and only if*

$$\lambda < \lambda_0 := (1 + w) \left(1 + w \max \left\{ \frac{a}{p}, \frac{1-a}{1-p} \right\} \right)^{-1}. \quad (35)$$

Proof. We can rewrite (34) in the following way

$$\max\{j_{-1}^{[\lambda C] - [a[\lambda C]]}, j_1^{[a[\lambda C]]}\} \leq \frac{C(w+1) - [\lambda C]}{w}.$$

Note that

$$\begin{aligned} &\mathbb{P} \left(j_1^{[a[\lambda C]]} \leq \frac{C(w+1) - [\lambda C]}{w} \right) \\ &= \mathbb{P} \left(\frac{j_1^{[a[\lambda C]]} - \frac{[a[\lambda C]]}{p}}{\sigma \sqrt{[a[\lambda C]]}} \leq \frac{1}{\sigma \sqrt{[a[\lambda C]]}} \left(\frac{C(w+1) - [\lambda C]}{w} - \frac{[a[\lambda C]]}{p} \right) \right) \end{aligned}$$

where $\sigma^2 := \frac{1-p}{p^2}$ (the variance of a random variable following the Geometric law with parameter p). The central limit theorem for i.i.d. random variables having finite variance

yields that $j_1^{[a\lambda C]} \leq \frac{C(w+1)-[\lambda C]}{w}$ with probability tending to one as $C \rightarrow \infty$ if and only if $(w+1-\lambda)/w - a\lambda/p > 0$, that is

$$\lambda < (1+w) \left(1 + \frac{aw}{p}\right)^{-1}. \quad (36)$$

In a similar way we can claim that

$$\mathbb{P} \left(j_{-1}^{[\lambda C] - [a\lambda C]} \leq \frac{C(w+1) - [\lambda C]}{w} \right) \rightarrow 1, \quad C \rightarrow \infty,$$

if and only if

$$\lambda < (1+w) \left(1 + \frac{(1-a)w}{1-p}\right)^{-1}. \quad (37)$$

Thus the statement of Lemma 1 is valid if and only if (36) and (37) are true simultaneously, i.e. relation (35) holds. \square

Thus the optimal boundary λ_0 is found.

5 Simulations

We will consider different levels of the total cost C and parameter γ of XOR-model. Note that now $\mathbb{P}(Y = 1) = \frac{\gamma}{2}$, i.e. $p = \frac{\gamma}{2}$. We employ two levels of parameter w . The value $w = 0$ corresponds to the extreme case when values of a response variable are obtained free of charge. This situation can be viewed as the limit one when the price of the measurement of Y is very low w.r.t. the price of X measurement. According to [22] the application of MDR method is reasonable for balanced datasets, i.e. with equal number of cases and controls. Therefore, in the case of the stratified samples we consider $a = 0.5$ and only even N in (23). For each combination of these parameters we generate D independent datasets for iMDR-EFE and for sMDR-EFE applications in order to evaluate performance of these methods by Monte Carlo experiments.

The values of parameters used in our simulations are given in Table 1. For relevant factors X_{k_1}, \dots, X_{k_r} we set $p_{k_i} = 0.5$, $i = 1, \dots, r$ (according to XOR-model) whereas for other (non-relevant) factors X_i the corresponding p_i are drawn independently from the uniform distribution $U(0.05, 0.5)$ to generate each dataset. For our datasets collections of p_i are drawn independently. Such simulation setting for p_i was proposed in [10]. The interval $[0.05, 0.5]$ for uniform law was taken since p_i is considered as MAF.

We take a Bernoulli variable $B(\gamma)$ having the success probability γ . Let $B(\gamma)$ and X be independent. Introduce $Y := 2B(\gamma)\mathbb{I}\{\sum_{v=1}^r X_{k_v} \bmod 2 = 1\} - 1$ where X_{k_1}, \dots, X_{k_r} are the relevant factors. Then one can verify that (28) holds. In such a way we generate independent vectors (X^j, Y^j) , $j = 1, 2, \dots$

Taking $w = 0.1$ we will assume that p is known in order to use it for determining $s'_{str}(C, w, \alpha)$ introduced in (23). For $w = 0$ the size of the sample $s'_{str}(C, 0, \alpha)$ is equal to C as in the case of independent observations. Then it is non-random and fixed before the experiment and we estimate $\mathbb{P}(Y = 1)$ by means of $Y^1, \dots, Y^{\tilde{C}}$. Here $\tilde{C} = \tilde{C}(C, a, \omega) = \max\{j_{-1}^{C-1}, j_1^{C_1}\}$, $C_1 = [aC]$ and $C_{-1} = C - [aC]$. In order to measure the method performance power we use TMR (true model rate) which is defined as

$$TMR := \frac{1}{D} \sum_{d=1}^D T_d$$

Parameter	Value
$\psi(y)$	$(\mathbb{P}(Y = y))^{-1}$
D	100
n	100
K	5
r	3
(k_1, \dots, k_r)	(2,3,5)
a	0.5
C	5 levels: $\{100, 200, 300, 400, 500\}$
w	2 levels: $\{0, 0.1\}$
γ	3 levels: $\{0.05, 0.1, 0.2\}$
α	0.05

Table 1: Simulation scheme parameters.

where

$$T_d = \begin{cases} 1, & \text{if all relevant factors are identified correctly in } d\text{-th dataset,} \\ 0, & \text{otherwise.} \end{cases}$$

Below the phrase “implementation of iMDR-EFE (sMDR-EFE) for sample” means that we select such collection of r factors which has the minimal \widehat{Err}_K .

The simulation procedure can be described in the following way.

- I. Fix some values γ and C from Table 1, other parameters except w are fixed as well.
- II. Calculate $s'_{str}(C, w, \alpha)$ where $w = 0.1$ and $\alpha = 0.5$ assuming that $p = \mathbb{P}(Y = 1)$ is known (and equals $\gamma/2$).
- III. Perform D times
 1. generation of independent $p_i \sim U(0.05, 0.5)$ for non-relevant factors;
 2. generation of independent X^j , having law introduced by (27), and corresponding Y^j until we have 3 samples:
 - (a) sample ξ_C which consists of C independent observations;
 - (b) stratified sample ζ_N which consists of $N = s'_{str}(C, w, \alpha)$ observations (for $w = 0.1$);
 - (c) stratified sample ζ_C which consists of C observations (for $w = 0$). While generating observations (X^j, Y^j) in order to form ζ_C we estimate p (and write “ p is estimated by means of observations $Y^1, \dots, Y^{\tilde{C}}$ ”);
 3. implementation of iMDR-EFE for ξ_C , p is unknown ($\psi(y)$ as well);
 4. implementation of iMDR-EFE for ξ_C , p is known therefore $\psi(y)$ is known and

$$\widehat{\psi}(y, \xi_C(\overline{S_k(C)})) = 1/\mathbb{P}(Y = y),$$

here we use the notation $\widehat{\psi}(y, \xi_C(\overline{S_k(C)}))$ of [4] for i.i.d. observations;

5. implementation of sMDR-EFE for ζ_N where $N = s'_{str}(C, w, \alpha)$, p is known therefore $\hat{P}_N = (1 - p, p)$ and

$$\hat{\psi}(y, \zeta_N(\overline{S_k(N)}), \hat{P}_N) = 1/P(Y = y);$$

6. implementation of sMDR-EFE for ζ_C , p is unknown and estimated by means of observations $Y^1, \dots, Y^{\tilde{C}}$;

7. implementation of sMDR-EFE for ζ_C , p is known thus $\hat{P}_C = (1 - p, p)$ and

$$\hat{\psi}(y, \zeta_C(\overline{S_k(C)}), \hat{P}_C) = 1/P(Y = y);$$

8. assignment of T_d to every implementation of the method under consideration.

IV. Evaluate TMR for each approach.

The results of our simulations are shown in Figures 1 and 2. Thus for the fixed total cost C a stratified sampling gives better results than independent one. Despite of the fact that $s_{ind}(C, \alpha) > s'_{str}(C, w, \alpha)$ (when $w > 0$, see Figure 2), sMDR-EFE demonstrates better performance than iMDR-EFE in all 3 models for each value of C and different scenarios concerning p . Moreover, small w permits to include more observations into the stratified sample. Figure 1 suggests that it leads to better performance of the sMDR-EFE method.

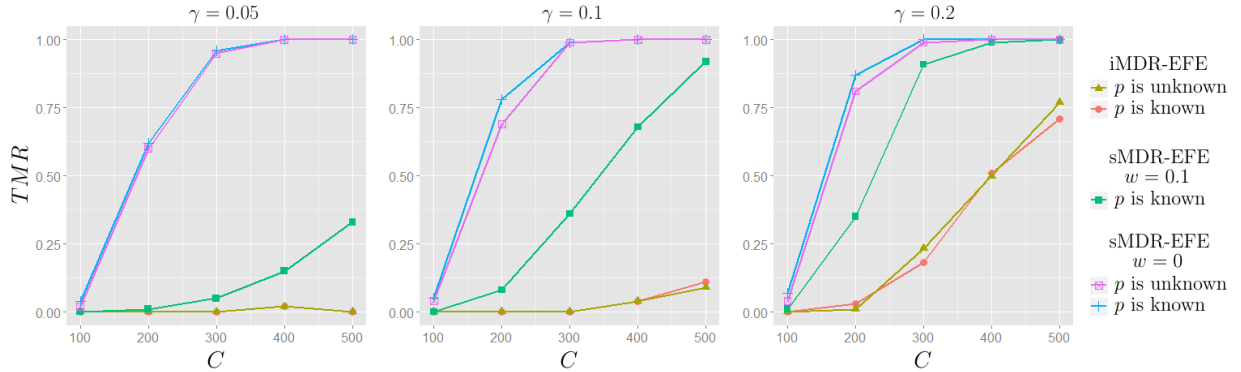


Figure 1: Performance (TMR) of iMDR-EFE and sMDR-EFE for different levels of $\gamma = 2p$.

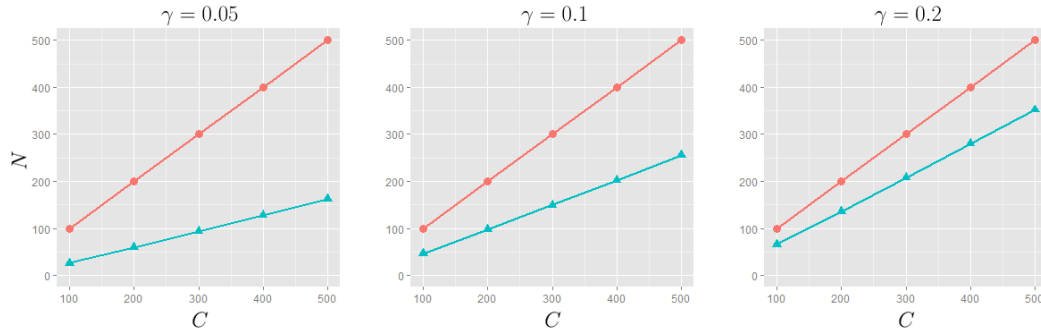


Figure 2: Sizes of the samples depending on the cost C for different levels of $\gamma = 2p$. The red line corresponds to iMDR-EFE and sMDR-EFE with $w = 0$, the blue line corresponds to sMDR-EFE with $w = 0.1$.

Let $a = 0.5$ and $w = 0.1$. Then, for γ taking values 0.05, 0.1 and 0.2 ($p = \gamma/2$) Lemma 4 yields that the corresponding values of λ_0 are equal to 0.366(6), 0.55 and 0.733(3),

respectively. For any $a \in (0, 1)$, all $p \in (0, 1)$ and $w = 0$ we get $\lambda_0 = 1$. Clearly if w is close to 0 then λ_0 is close to 1.

We note that there are various possibilities to take into account the cost of experiments with random and nonrandom number of observations. This interesting problem will be considered separately.

6 Acknowledgements

The work of A.V.Bulinski is supported by the Russian Science Foundation under grant 14-21-00162 and performed at the Steklov Mathematical Institute of Russian Academy of Sciences. The problems setting belongs to A.V.Bulinski, all theoretical results are established jointly with A.A.Kozhevin. Computer simulations are carried out by A.A.Kozhevin.

References

- [1] S.E.Ahmed. Penalty, Shrinkage and Pretest Strategies. Variable Selection and Estimation. Springer, Cham, 2014.
- [2] V.Bolón-Canedo, N.Sanchez-Marono and A.Alonso-Betanzos. Feature Selection for High-Dimensional Data. Springer, Cham, 2015.
- [3] P.Bühlmann, S.van de Geer. Statistics for High-Dimensional Data. Methods, Theory and Applications. Springer, Heidelberg, 2011.
- [4] A. Bulinski. On foundation of the dimensionality reduction method for explanatory variables. Journal of Mathematical Sciences, v. 199, No. 2, 113-122 (2014).
- [5] A.Bulinski. Central limit theorem related to MDR-method. In: Asymptotic Laws and Methods in Stochastics. A volume in Hounor of Miklos Csorgo. Fields Institute Communications, v. 76, 113-128. Springer, New York, 2015.
- [6] A.Bulinski. Some statistical methods in genetics. In: V.Schmidt (Ed.). Stochastic Geometry, Spatial Statistics and Random Fields. Lecture Notes in Mathematics, v. 2120, 293-320. Springer-Verlag, Berlin, 2014.
- [7] A.Bulinski, O.Butkovsky, V.Sadovnichy, A.Shashkin, P.Yaskov, A.Balatskiy, L.Samokhodskaya and V.Tkachuk. Statistical methods of SNP data analysis and applications. Open Journal of Statistics, v. 2, No 1, 73-87 (2012).
- [8] A.Bulinski, A.Rakitko. Simulation and analytical approach to the identification of significant factors. Commun. in Statistics. Part B: Simulation and Computation, v. 44, 1-23 (2015).
- [9] A.Bulinski, A.Rakitko. MDR method for nonbinary response variable. J. of Multivariate Analysis, v. 135, 25-42 (2015).
- [10] A.Deelman, C.Ambroise and P.Neuviat. Performance of a blockwise approach in variable selection using linkage disequilibrium information. BMC Bioinformatics 16:148 (2015).

- [11] K-A.Do, Z.S.Qin and M.Vannucci (Eds.). *Advances in Statistical Bioinformatics. Models and Integrative Inference for High-Throughput Data*. Cambridge University Press, Cambridge, 2013.
- [12] D.Gola, J.M.M.John, K. van Steen and R.König. A roadmap to multifactor dimensionality reduction methods. *Briefings in Bioinformatics*, June 24, 1-16 (2015).
- [13] G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*, Springer Science + Business Media, New York, 2013.
- [14] I.Koch. *Analysis of Multivariate and High-Dimensional Data*. Cambridge University Press, Cambridge, 2014.
- [15] J-M.Marin, C.Robert. *Bayessian Essentials with R*. Springer Science + Business Media, New York, 2014.
- [16] J.H.Moore, S.M.Williams (Eds.). *Epistasis: Methods and Protocols*. *Methods in Molecular Biology*. v. 1253. Springer Science + Business Media, New York, 2015.
- [17] J. Park, Independent rule in classification of multi- variate Binary Data, *J. of Multivariate Analysis*, v. 100, No. 10, 2270-2286 (2009).
- [18] M. D. Ritchie, L. W. Hahn, N. Roodi, R. Bailey, W. D. Dupont, F. F. Parl and J.H. Moore, Multifactor dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, *Amer. J. Human Genetics*, v. 69, 139-147 (2001).
- [19] G.Ritter. *Robust Cluster Analysis and Variable Selection*. CRC Press, Boca Raton, 2015.
- [20] J. Shang, J. Zhang, Y. Sun, D. Liu, D. Ye and Y. Yin, Performance analysis of novel methods for detecting epistasis *BMC Bioinformatics* 12:475 (2011).
- [21] R. L. Taylor and T.-C. Hu. Strong laws of large numbers for arrays of row-wise independent random elements, *Int. J. Math. Math. Sci.*, v. 10, 805-814 (1987).
- [22] D. Velez, B. White, A. Motsinger, W. Bush, M. Ritchie, S. Williams, and J. Moore. Balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction, *Genetic Epidemiology*, v. 31, 306-315 (2007).
- [23] S.J.Winham, A.J.Slater and A.A.Motsinger-Reif. A comparison of internal validation techniques for multifactor dimensionality reduction. *BMC Bioinformatics*, 11:394 (2010).